DOCUMENT RESUME

ED 190 603 TM 800 395

AUTHOR Haladyna, Tom; Roid, Gale
TITLE A Comparison of Decision-Making Methods for Criterion-Referenced Tests.
PUB DATE Apr 80
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Bayesian Statistics; *Classification; Comparative Analysis; Criterion Referenced Tests; *Cutting Scores; *Decision Making; Elementary Education; Higher Education; Latent Trait Theory; Mastery Tests; *Mathematical Models; *Reliability; Test Results; *Test Theory
IDENTIFIERS Binomial Coefficients; Classical Test Theory; Generalizability Theory; *Test Length

ABSTRACT

The problems associated with misclassifying students when pass-fail decisions are based on test scores are discussed. One protection against misclassification is to set a confidence interval around the cutting score. Those whose scores fall above the interval are passed; those whose scores fall below the interval are failed; and those whose scores fall within the interval are retained for brief remedial instruction. Six methods which may be used to set the confidence level include: (1) Novick and Jackson's Bayesian method, which uses a computer program for statistical analysis, Computer-Assisted Data Analysis; (2) Millman's binomial method; (3) Shaycroft's variation of the binomial method; (4) Lord and Novick's classical approach; (5) Cronbach et. al.'s generalizability theory; and (6) Rasch's latent trait model. These six methods were empirically compared using criterion-referenced test score data from 80-item and 100-item tests of dental anatomy, as well as 68-item elementary school achievement tests. The interaction of the methods with passing standards and with test length was also investigated. While the results indicated promise for Bayesian and item-response (Rasch) theories, no single procedure proved to be more consistent in minimizing classification errors. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190603

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Comparison of Decision-Making Methods for Criterion-Referenced Tests[1]

Tom Haladyna
Teaching Research Division
Oregon State System of Higher Education
Monmouth, Oregon 97361

Gale Roid
Oregon Department of Human Resources

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Haladyna

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800395

[1] This research was supported through a grant from the National Institute of Education. Opinions expressed in this paper are those of the authors and do not represent the official National Institute of Education position or policy.

A Comparison of Decision-Making Methods for Criterion-Referenced Tests

Tom Haladyna
Gale Roid

In most forms of systematic instruction, objectives are used to guide both instruction and testing, and criterion-referenced tests are used to measure student progress in sequential units (Bloom, 1968; Hambleton, 1974; Keller, 1968). Students whose scores fall at or near a criterion level are in danger of being misclassified due to errors of measurement. More importantly, many of these students are in need of only minor remediation, as opposed to those failing far below this criterion level who are in need of reassignment to more appropriate instruction or reinstruction. A number of statistical strategies are available for assigning students to pass or fail categories, and it is the purpose in this study to compare these approaches.

The problem of misclassifying examinees is considerably more widespread than with the confines of systematic instruction. Most professional training institutions use exams to certify their graduates, and professions such as law and medicine provide similar exams for licensure. More recently, many states have mandated the measurement of high school students' competencies and the certification of these competencies as part of graduation requirements. Thus, the problem of minimizing errors of classification is found in a number of situations both within and outside of systematic instruction, and it is a problem which has received considerable attention from measurement theoreticians (e.g., Brennan, 1978; Brennan & Kane, 1977; Hambleton & Novick, 1973; Millman, 1974b).

Although the term "criterion-referenced test" has had a number of different interpretations (Glass, 1978; Hambleton, Swaminathan, Algina & Coulson, 1978; Popham, 1978), it is clear that the intent behind this approach to testing is to (a) specify the intent of instruction in terms of instructional objectives or domain specifications, (b) develop a universe of test items representing this instructional intent, and (c) take random samples of items to create test forms. The underlying assumptions are that an item universe is defined and that items are keyed to objectives and instruction itself, and this integration of teaching and testing is the hallmark of systematic instruction (Shoemaker, 1975).

Working within the framework of these test construction procedures, we can conceptualize any student's score representing an objective, a group of objectives or a domain in terms of true and observed scores, as is done in classical test theory and generalizability theory. The true score is the percentage of items correctly answered from the administration of all items in the domain. The observed score is the result obtained for any student from any sample of items from that domain. As Hambleton and Novick (1973), among others, have observed, one is prone to make two kinds of errors: (1) assigning a mastery student to non-mastery status or (2) assigning a non-mastery student to mastery status.

A more reasonable instructional strategy was proposed by Millman (1974a) when he suggested that an uncertainty band be constructed around the passing standard. Those having scores found in the uncertainty band would be assigned to remedial instruction, while those in the lower end of the scale would receive reinstruction or reassignment to other instruction of a more elementary or differing nature (see Figure 1).

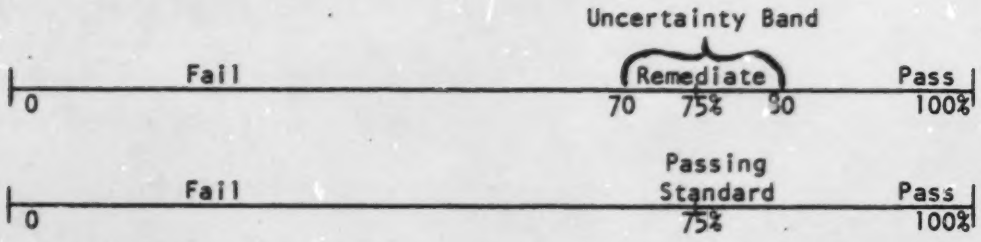

Figure 1. Two strategies to decision making in systematic instruction.

- - - - - - - -

Since the earlier statement of this problem by Emrick (1971) and by Hambleton et al. (1973), there has been increased interest in this problem, and there have been many proposals, representing widely divergent theoretical viewpoints, for how best to assign students to various categories. To date, these approaches have not been empirically contrasted, and there is little information regarding the effectiveness of each model in helping to decide which students pass, fail or are held for remedial instruction. The balance of this paper contains two parts. The first is intended to describe five theoretical approaches to decision making and related research, and the second is devoted to a report of the empirical comparison of these approaches.

Decision-Making Strategies

Bayesian

The work of Novick and Jackson (1974) currently forms the theoretical basis for the application of Bayesian methods to the problem of decision making. Perhaps the most pertinent applications of their work can be found in the seminal paper by Hambleton et al. (1973) and later in the studies of

Swaminathan, Hambleton, and Algina (1974) and Hambleton, Hutten, and Swaminathan (1976). The Bayesian approach incorporates prior information into the model which is purported to reduce uncertainty and thereby decrease errors of misclassification. Unlike most other approaches, these procedures employ transformations of scores to obtain more precise estimates. One accomplishes this by a series of statistical procedures which can best be done by a computer program, Computer-Assisted Data Analysis (CADA), made available from the University of Iowa.

Hambleton et al. (1976) studied the accuracy of decision making with simulated data using several classical and Bayesian methods. Variables of interest in this study were test length, sample size, and homogeneity of the true mastery score distribution. Findings revealed a slight superiority for Bayesian methods for estimating true scores. Sample sizes and test length did not seem to be importantly related to the criterion measures of goodness of fit that were used in this study.

Binomial

In this approach (Millman, 1974a), the uncertainty band is constructed using the parameters of domain size, test length, and passing standard. Haladyna (Note 1) compared this binomial method to a classical procedure and found that both led to rather large uncertainty bands in a wide variety of settings. Both approaches appeared to be lacking in terms of utility. With the binomial approach, uncertainty was a function of the size of the domain and the passing standard. Despite the temptation to manipulate the passing standard to minimize uncertainty, there seems little justification to tamper with the passing standard which is determined by a different set of procedures that have nothing to do with reliability (see Meskauskas, 1976;

Millman, 1973), but everything to do with validity (see Glass, 1978; Messick, 1975).

A second approach, suggested by Shaycroft (Note 2; Note 3), employs the same variables as Millman's, but includes the parameter for the number of options for each item and does not involve a passing standard. Tables are provided for decision making which do not reflect, in any way, any other characteristic of the actual test or the results of the test, such as the mean or standard deviation, the quality of the items, or internal consistency of the items.

Classical

From a strictly classical approach to the problem (Lord & Novick, 1968; Nunnally, 1967), the uncertainty band could be constructed using the standard error of measurement, which is analogous to the standard error of estimate in the Bayesian approach and the uncertainty band in the binomial approach. Despite beliefs by many that classical reliability concepts do not apply to CR tests, studies by Bernknopf and Bashaw (Note 4), Epstein and Knerr (Note 5), Haladyna (Note 6) and Moyer and Fishbein (Note 7) present arguments for using KR-20 with combined pretest and posttest data.

As indicated earlier, the study by Hambleton et al. (1976) using simulated data and the study by Haladyna (Note 1) using actual test data provided some information about the relative merits of Bayesian, binomial and classical approaches. If anything conclusive can be drawn from these two studies, it appears that the binomial and classical approaches may not offer very much precision in decision making. However, if one considers feasibility and cost as factors in decision making, then the binomial and classical approaches are far ahead.

Generalizability Theory

The work of Cronbach and his colleagues (Cronbach, Gleser, Nanda & Rajaratnam, 1972) has motivated the application of generalizability theory to decision making in systematic instruction. The extent of this application is well represented in the work of Brennan (1978) and Brennan and Kane (1977). Generalizability theory is based on the idea that tests are samples from well defined universes of items. Therefore, the theory is compatible with the problem of decision making within systematic instruction. More importantly, the theory liberalizes classical theory by departing from the classical assumption that errors of measurement are unidimensional. In contrast to classical theory, generalizability theory takes into account various sources of errors and uses analysis of variance procedures to estimate reliability and measurement error.

Empirical studies where the generalizability theory is compared with others have not been performed. However, the simplicity of the procedures, coupled with some of the attractive features of generalizability theory, make such empirical comparisons most promising.

Latent-Trait Theory

The last of the approaches to decision making to be discussed involves the one-parameter latent-trait (Rasch) model advocated by Wright (1967). The Rasch model is the simplest of a family of latent-trait models (see Hambleton & Cook, 1977, for a discussion of the models). With the Rasch model, a test construction strategy is used which minimizes errors of estimate of true scores (analogous to errors of measurement). By selecting items with moderate difficulty for each examinee, standard errors are smallest and classification errors should be minimized. There is no research yet

reported which attests to the relative accuracy of decision making with the Rasch model in contrast to other approaches. What is particularly appealing about this approach is that the selection of items governs the size of errors of measurement and provides the precision required to minimize these errors. Latent-trait theory, unlike others, may provide a more effective solution to the problem of decision making without encountering many of the difficulties reported in theoretical and empirical studies by Haladyna (Note 1), Haladyna and Roid (Note 8), and Hambleton et al. (1976).

Other Approaches

There are a number of other approaches to CR reliability, which for lack of space cannot be reviewed here. More importantly, these omitted approaches have emphasized a coefficient of reliability, or an index of correct classification that does not permit an estimation of measurement error vital for decision making. Furthermore, there have been an overabundance of CR reliability indexes. The title of Frederick's paper (Note 9) captures the essence of the problem here: "A(nother) model for the reliability of criterion-referenced tests."

Also omitted from this review is the work of Huynh (1976a, 1976b, Note 10, Note 11) with the beta-binomial model and other recent theoretical developments. Although quite promising, methodologies for empirically testing these models have not yet been fully developed. This, however, should not rule out their use in future studies.

In summary, there has been a recent and prolific production of studies on CR test reliability as it applies to decision making in systematic instruction. Most of the work has been theoretical with only two studies reported that involve test data, one simulated, the other real. Clearly, there is a

need to examine each of these theories using achievement test data in the context of systematic instruction. Further, it is important to systematically vary conditions of data, such as test length and criterion levels, to adequately test these rivaling approaches. As noted earlier, the second objective in this paper was to determine which of the theoretical models presented minimizes errors of misclassification. Specifically, the research questions of the study were:

1. Which procedure most accurately classifies examinees?

2. What are the observable strengths and weaknesses, advantages and disadvantages, of each approach under the conditions of varying data sets, passing standards and test lengths?

Method

Design of the Study

A true score in this study was any student's score on the entire item universe. In each of these cases, the item universes were finite, ranging from 68 to 100 items. Observed scores were determined by taking a random sample of items of a specified length and scoring that sample. All results were reported in the metric of percentage correct, which is an unbiased estimate of the universe score (Harris, Pearlman & Wilcox, 1977; Millman, 1974a).

Given any passing standard, it is possible to assign any student to pass or fail based on their true score. Given a student's observed score and an uncertainty band derived from one of the models discussed earlier, and it is also possible to assign the student to one of three states, pass, uncertainty, fail. With these pieces of information it is possible to ascertain the

per cent of correct classification. The chart below captures the essence of six possible conditions that may occur for any examinee:

	Procedure Classifies Student as		
	Definite Fail	Uncertain	Definite Pass
True Pass	Incorrect (ICP)	Remedial (UCP)	Correct (CCP)
True Fail	Correct (CCF)	Remedial (UCF)	Incorrect (ICF)

The dependent measures of the study were three possible conditions. The first was the percentage of incorrect classifications (ICF + ICP), the second was the percentage of persons classificed as uncertain who will receive remedial instruction and retesting (UCF + UCP), and the third was the percentage of correct classifications (CCF + CCP). It is important to note that the first two measures represent errors, the former being most serious. In the latter, anyone classified as "uncertain" is truly a pass or a fail, but is close enough to the passing standard to warrant remedial instruction. Wide uncertainty bands will reduce the most serious kind of errors of misclassification (ICP and ICF) at the expense of increasing the second type of errors of misclassification (UCP and UCF). Larger uncertainty bands will decrease the percentage of correct classifications as well.

The independent variables of these studies were (a) four data sets which varied in terms of the magnitude of the differences between pretest and posttest means, (b) five theoretical approaches with one variation in the binomial approach which makes a total of six different methods for setting uncertainty bands, (c) six criterion levels, ranging from .60 to .90 in increments of .05, and (d) test length, ranging from 10 to 40 in increments of 10.

To protect against chance fluctuations due to the sampling, with each test length three forms were produced. This increased the number of observations upon which the analysis was conducted.

Sources of Data

Test data which provided true scores or simulated true scores were taken from the following sources:

1. A 100-item domain covering objectives related to a five-volume text on Dental Anatomy (Roid, 1977). Approximately 250 students were given the entire 100-item set before and after instruction.

2. An 80-item domain was accepted as representing an item universe covering one of the five volumes of the text mentioned above. In this instance, about 200 students were administered the test both before and after instruction.

3. Two elementary-grade level, objective-based achievement tests from the Beaverton (Oregon) Public Schools.

Descriptive statistics for these data are given in Table 1.

Table 1

Domain Size, Sample Size, Means and Standard Deviations[1] for Test Data Used in the Analyses

Data Set	Domain Size	Pretest n	Pretest mean	Pretest s.d.	Posttest n	Posttest mean	Posttest s.d.	Sensitivity to Instruction
1. Dental Anatomy	80	152	34.9	7.6	136	91.2	12.5	56.3
2. Dental Anatomy	100	256	32.9	13.7	254	74.8	17.9	41.9
3. Elem. 1	68	291	25.4	9.4	236	51.3	19.0	25.9
4. Elem. 2	68	306	28.5	10.7	326	46.9	15.1	18.4

[1] All means and standard deviations are reported in percentages.

Analysis of Data

Estimating error. Comparable standard errors were computed for each of the six methods using recommended procedures. Each of these procedures leads to the estimation of a standard error for the creation of an uncertainty band where students are classified as (a) definite pass, (b) uncertain, (c) definite fail.

The binomial procedure involves simple calculations using a formula provided by Millman (1974b, p. 18). The procedure requires simple arithmetic involving the parameters of domain size, passing standard and test length. Since a domain can be finite or infinite, two variations of this model are possible. For purposes of the study standard errors for both finite and infinite domains were computed.

The Bayesian procedure was implemented using the Computer Assisted Data Analysis (CADA) interactive computer program provided by the University of Iowa. The program utilizes information on pretests to establish parameters which are used in a posterior analysis. This program was based on the work of Novick, Lewis and Jackson (1973). These procedures lead to estimates of the probability that a proporation of correct responses exceeds a given criterion level. Based on these estimations, uncertainty bands were created and used.

The classical procedure is implemented by first estimating reliability using the conventional KR-20 internal consistency estimate and then the familiar formula for the standard error of measurement.

Generalizability theory allows for the estimation of reliability as a function of the criterion level rather than the mean of the group of examinees being studied (Brennan & Kane, 1977). This is similar, but also

in contrast, to the procedures earlier recommended by Livingston (1972). Brennan and Kane (1977) provide a thorough discussion of the similarities and differences. A computer program was created which estimates reliability and the standard error (using the procedures recommended by Brennan & Kane).

Criterion measure. The basis for evaluating these six methods for setting uncertainty bands was the set of three indexes which are simply the proportions of total examinees who were (a) incorrectly classified, (b) classified as uncertain and (c) correctly classified. To compute any of these indexes one needs the true score of each examinee, the criterion level for the test, and the test result. True scores (or simulated true scores) were available for each data set, and true scores were observed for each examinee across each of the varying passing standards and test lengths.

An analysis of variance was performed for each of these three dependent variables. The design for this analysis was a 4 x 6 x 7 x 4, which included four data sources, six uncertainty bands, seven passing standards, and four test lengths. The analyses were restricted to main effects and first order interactions. All other sources of variance were considered as residual in the analysis.

With large sample sizes, small and often trivial differences are statistically significant. Thus, a two-stage decision process was used. First, alpha was set at .001. Second, when results were statistically significant, a test of practical significance was used where the effect size (the difference between two units of analysis) had to exceed a .50 standard deviation for the particular level in the experimental design. These procedures thus insured that any finding would have to be practically, in addition to statistically, significant.

Results and Discussion

The percentage of students who were true pass and true fail were tabulated for each category of observed test scores--fail, uncertain, and pass. These results are given in Table 2. Roughly 57% of all student scores were classified as true fail, while about 43% were classified as true pass. The six decision-making methods across all data sets, passing standards and test lengths resulted in only 1.64% misclassifications and 23.19% classified as uncertain. As noted previously, the first type of error is most serious, while the second type of error forces students into remedial efforts, even though many of these students classified as uncertain are true passes. Although more true fails were classified as uncertain when contrasted with true passes (12.35% to 10.84%), this difference was not of a significant magnitude. Thus, all procedures appeared to not show a systematic bias for misclassifying students.

Table 2

Summary of Classifications Across All Conditions

Classified as: as:	True Fail	True Pass	Total
Fail	CCF 44.14%	ICP 00.70%	44.84%
Uncertain	UCF 12.35%	UCP 10.84%	23.19%
Pass	ICF 00.94%	CCP 31.03%	31.97%
Total	57.43%	42.57%	100.00%

The results of the analysis of variance for each of the dependent measures and all independent variables is given in Table 3. As shown there, most main effects and interactions were statistically significant at .001. When the criterion of practical significance was applied, 16 of 30 results were practically significant. Of these 16 significant findings, only three involved the methods of assigning students. The first part of this section deals with the accuracy of decision making as a function of the six methods studied, while the second part deals with other findings of interest.

Methods for Assigning Students

The main purpose of this study was to compare the effectiveness of six approaches to statistical decision making to reduce classification errors in systematic instruction. The results of the analysis of variance revealed that the approach to decision making did not make a practically significant difference in the proportion of incorrect classifications.

The percentages for each of the three categories by method were:

	Incorrect	Uncertain	Correct
1. Bayesian Decision Theoretic	1.23	22.38	76.39
2. Binomial (Finite Domain)	2.26	20.01	77.72
3. Binomial (Infinite Domain)	1.57	24.31	74.12
4. Classical	1.95	21.15	76.90
5. Generalizability	1.59	23.57	74.84
6. Rasch	1.21	27.71	71.09
Total (all approaches combined)	1.64	23.19	75.18

Table 3

Results of Analyses of Variance for Each of the Three Dependent Measures of the Study

Source of Variation	df	Type of Classification: Incorrect F	Incorrect p < .001	Incorrect Pract. Signif.[1]	Uncertainty F	Uncertainty p < .001	Uncertainty Pract. Signif.	Correct F	Correct p < .001	Correct Pract. Signif.
Data Set (A)	3	56.2	yes	yes	607.6	yes	yes	151.5	yes	yes
Theory (B)	5	6.7	yes	no	20.0	yes	no	29.4	yes	no
Passing Standard (C)	6	5.9	yes	no	25.2	yes	no	54.5	yes	no
Test Length (D)	3	49.6	yes	yes	157.9	yes	yes	393.8	yes	no
A x B	15	4.2	yes	no	5.0	yes	no	7.5	yes	no
A x C	18	41.1	yes	yes	89.2	yes	yes	198.1	yes	yes
A x D	15	11.3	yes	yes	8.3	yes	yes	17.9	yes	yes
B x C	30	2.4	yes	no	2.2	yes	yes	3.2	yes	yes
B x D	15	3.1	yes	yes	0.9	no	--	2.5	no	--
C x D	18	4.6	yes	yes	1.7	no	--	2.3	yes	yes
Total df	1893									

[1] Practical significance was determined by noting if the difference between any extreme mean in a level or simple effect and the grand mean for that level or simple effect exceeded .50 standard deviations. The selection of .50 as a standard for this effect size was an arbitrary choice with the intention to impose a more stringent criterion than statistical significance.

The Bayesian decision-theoretic procedures and the use of the Rasch model provided a slightly lower percentage of incorrect classification. However, the Rasch model reduces error at the expense of placing more students in the uncertainty range. The binomial approach, where the domain is conceptualized as finite, yielded the highest error rate, 2.26, but also has the highest percentage of correct classifications and the lowest percentage of students classified as uncertain.

Interaction of these methods with passing standards. The cell means for this interaction including each of the dependent variables is presented in Table 4. Only the second and third dependent variables, uncertainty and correct classification, yielded practically significant interactions and these were minor effects indeed.

With respect to correct classifications, most procedures give high classification accuracy at higher passing standards with minor variations in this tendency. The number of students found in the uncertainty range fluctuated without pattern from method to method with lower percentages existing when the passing standard was highest. Incorrect classifications were found with the same incidence in virtually all conditions with lower rates at the high and low passing standards.

Interaction of these methods with test length. The cell means for this interaction for each of the dependent variables are presented in Table 5. Only the percentage of incorrect classification yielded a practically significant interaction. The main cause for this interaction seems to be the higher percentage of error found with the 10-item test forms. As shown in Figure 1, the Bayesian approach appears to be the least error-prone of other methods at the 10-item test length, while the Rasch model yields the

Table 4

Mean Percentages of Classification for the Interaction of Decision-Making Method with Passing Standard

Method	Incorrect Classification						
	.60	.65	.70	.75	.80	.85	.95
Bayesian	0.9	0.8	1.3	1.8	1.9	1.6	0.3
Binomial (Finite)	1.8	2.7	2.0	3.6	2.3	1.6	1.8
Binomial (Infinite)	1.3	0.8	1.4	2.7	1.7	1.4	1.8
Classical	1.7	2.5	1.9	2.9	2.0	1.6	1.0
Generalizability	1.4	1.7	1.5	2.5	1.6	1.4	1.0
Rasch	1.2	0.9	1.2	1.7	1.4	1.2	1.0

Method	Uncertainty Classification						
	.60	.65	.70	.75	.80	.85	.95
Bayesian	24.7	26.8	27.4	25.3	21.5	15.2	15.6
Binomial (Finite)	20.6	18.3	23.5	21.3	22.7	19.5	14.2
Binomial (Infinite)	25.1	31.6	28.8	22.6	26.7	21.2	14.2*
Classical	20.4	18.9	24.6	21.9	23.8	19.5	18.8
Generalizability	22.8	23.9	26.3	25.4	25.4	21.8	19.5
Rasch	27.0	30.0	30.4	30.0	29.3	27.2	20.0

Method	Correct Classification						
	.60	.65	.70	.75	.80	.85	.90
Bayesian	74.4	72.3	71.3	72.9	76.6	83.2	84.1
Binomial (Finite)	77.6	79.0	74.4	75.0	75.0	78.9	84.1
Binomial (Infinite)	73.7	67.6	69.8	74.7	71.6	77.4	84.1*
Classical	77.8	78.5	73.5	75.2	74.2	78.9	80.2
Generalizability	75.8	74.4	72.2	72.2	73.0	76.8	79.6
Rasch	71.8	69.2	68.4	68.4	69.3	71.6	79.0

* indicates practical significance $\left(\frac{\bar{X} - \bar{\bar{X}}}{s} > .50\right)$

Table 5

Mean Percentages of Classification for the Interaction of Decision-Making Methods with Test Lengths

Method	Incorrect Classification 10	20	30	40
Bayesian	1.8*	1.6	1.0	0.6
Binomial (Finite)	3.6*	2.8	1.4	1.2
Binomial (Infinite)	3.1*	1.7	1.0	0.6
Classical	3.4*	2.5	1.1	0.8
Generalizability	3.0*	1.7	1.0	0.7
Rasch	2.5*	1.3	0.7	0.4

Method	Uncertainty Classification 10	20	30	40
Bayesian	29.9	22.6	19.5	17.6
Binomial (Finite)	29.0	19.3	18.8	12.9
Binomial (Infinite)	32.2	25.9	21.0	18.2
Classical	29.6	19.7	19.2	16.0
Generalizability	32.1	24.4	20.6	17.2
Rasch	37.0	29.5	23.7	20.7

Method	Correct Classification 10	20	30	40
Bayesian	68.4	75.9	79.5	81.8
Binomial (Finite)	67.4	77.8	79.8	85.9
Binomial (Infinite)	64.8	72.5	78.0	81.2
Classical	67.0	77.8	79.6	83.2
Generalizability	64.9	73.9	78.3	82.2
Rasch	60.5	69.3	75.6	79.0

* Indicates practical significance ($\frac{\overline{X} - \overline{\overline{X}}}{s} > .50$)

highest degree of precision particularly at the longest test length of 40 items. The other four approaches are distinctly homogeneous with respect to error rates.

Other Findings

Interaction of data sets with test lengths. As shown in Table 2, the main effects of data set and test length, and the interaction of these factors, were statistically and practically significant. The cell and marginal means for this interaction are given in Table 6. As shown there, data set one had an error rate of 3.2%, much higher than any of the other three. In addition, this data set had the highest number of students in the uncertainty band and the lowest number of students correctly classified. Error rates were smallest for data sets three and four, where most student scores were at or below the passing standards for the most part. Where the errors were highest, in data set one, most student scores fell in the range of passing standards or above.

Test length is strongly related to precision, as would be predicted from any theory of test scores. Longer tests did indeed reduce incorrect classifications and uncertainty classifications. What is most significant from these findings is that the rate at which errors occur was a function of data set. That is, there was a strong interaction between data sets and test lengths for each of the dependent variables of the study. What these results serve to illustrate is that the distribution of student scores strongly influences the magnitudes and types of errors and correct classifications.

Interaction of data set and passing standard. As with the interaction of data set with test length, a similar interaction existed between data set and passing standard. These results appear in Table 7. As shown there, errors

Table 6

Mean Percentages of Classification for the Main Effects and Interaction of Data Sets with Test Lengths

Data Set	Incorrect Classification				
	10	20	30	40	Total
One	5.3*	3.2	2.6	1.7	3.2*
Two	1.9*	1.0	0.5	0.3	1.0
Three	2.8*	1.6	0.5	0.4	1.3
Four	1.7	1.8	0.5	0.3*	1.1
Total	2.9	1.9	1.0	0.7	1.6

Data Set	Uncertainty Classification				
	10	20	30	40	Total
One	49.6*	41.0	38.2	33.4	40.5*
Two	15.5	15.0	11.5	8.4	12.6*
Three	30.5*	16.0	12.6	11.4	17.6
Four	30.8*	22.2	19.6	15.2	22.0
Total	31.6	23.6	20.5	17.1	23.2

Data Set	Correct Classification				
	10	20	30	40	Total
One	45.2*	55.8	59.2	65.0	56.3*
Two	82.6	84.0	87.9	91.2	81.0
Three	66.6*	82.4	86.9	88.1	77.0
Four	67.5*	76.0	79.9	84.5	86.4*
Total	65.5*	74.5	78.5	82.2	75.2

* Indicates practical significance $\left(\frac{\bar{X} - \bar{\bar{X}}}{s} > .50\right)$

Table 7

Mean Percentages of Classification for the Interaction of Data Sets with Passing Standards

Data Set	Incorrect Classification						
	.60	.65	.70	.75	.80	.85	.90
One	1.3*	2.4	2.8	6.2*	4.8	3.5	1.4
Two	0.0*	0.1	0.3	1.0	1.1	1.5	2.7*
Three	2.5*	2.1	1.8	1.7	0.5	0.4	0.2*
Four	1.6	1.7	1.3	1.3	0.8	0.5	0.2*

Data Set	Uncertainty Classification						
	.60	.65	.70	.75	.80	.85	.90
One	24.5*	35.4	49.8*	50.2*	54.8*	42.6	26.5*
Two	1.4*	3.7*	6.0	8.6	15.3	21.6*	31.8*
Three	33.9*	27.8*	22.3	16.1	12.0	7.6*	3.8*
Four	33.8*	32.8*	29.2*	22.8	17.6	11.3*	6.1*

Data Set	Correct Classification						
	.60	.65	.70	.75	.80	.85	.90
One	74.2	62.2	47.4	43.7	40.4	54.0	72.1
Two	98.5*	96.2*	93.7*	70.5	83.7	76.9*	65.5*
Three	63.5*	70.0*	75.8	82.2	87.4	92.1*	96.0*
Four	64.5*	65.5	69.5	75.9	81.6	88.2*	93.7*

* Indicates practical significance ($\frac{\bar{X} - \bar{\bar{X}}}{s} > .50$)

of classification and uncertainty classification were highest at the middle passing standards and lowest for the low and high passing standards. For the other data sets, errors rates depended heavily again on the distribution of test scores. For data set two, where student scores were relatively high, errors were highest when the passing standard was highest, as might be expected. For data sets three and four, errors were highest when passing standards were low, where most student scores congregated. Errors were lowest when the passing standards were highest, where the fewest number of students were observed. Again, these results indicate the powerful effect that the distribution of students' scores has on incorrect, uncertainty and correct classifications.

Interaction of passing standards and test lengths. The results for this interaction are shown in Table 8. As we might expect, errors are highest with short tests and lowest with long tests. The same holds true for uncertainty classification, and the reverse is true for correct classifications. With shorter tests, uncertainty bands are large which increases the number of students classified as uncertain. Small tests have sufficiently large errors of measurement to cause incorrect classifications.

The nature of the interaction of these two variables occurs with the dependent variables of correct and incorrect classifications. With incorrect classification, errors are infrequent for longer tests with greatest frequency for medial passing standards. In other words, fewer errors were detected when passing standards are very low or very high. With 10-item tests, error rates fluctuated erratically without pattern but were distinctly higher than for all other test lengths studied.

For the rate of correct classifications, the same general trends applied. Higher percentages of correct classification were made for tests 20 items or longer and with very high or very low passing standards.

Table 8

Mean Percentages of Classification for the Interaction of Passing Standards with Test Lengths

Test Length	Incorrect Classification .60	.65	.70	.75	.80	.85	.90
10	2.8*	3.2*	2.5*	3.9	2.2	3.0*	2.8*
20	1.6	1.8	1.8	3.3	2.8	1.5	0.7
30	0.7	0.8	1.0	1.9	1.4	0.8	0.6
40	0.5	0.5	0.8	1.1	1.0	0.5	0.4

Test Length	Uncertainty Classification .60	.65	.70	.75	.80	.85	.90
10	33.9	34.3	36.6	29.7	35.2	28.2	23.5
20	22.9	24.6	27.5	23.4	24.3	22.4	19.8
30	20.4	22.3	23.6	23.3	21.6	17.8	14.4
40	16.5	18.5	19.6	21.3	18.6	14.6	10.5

Test Length	Correct Classification .60	.65	.70	.75	.80	.85	.90
10	63.3*	62.5*	60.9*	66.4	62.7	68.8	73.7
20	75.5	73.6	70.7	73.3	72.9	76.1	79.5
30	78.9	76.9	75.3	74.8	77.0	81.3	85.0
40	83.0	80.9	79.5	77.6	80.4	84.9	89.1

* Indicates practical significance $\left(\frac{\bar{X} - \bar{\bar{X}}}{s} > .50\right)$

Conclusions

The objective of this study was to empirically study five decision making methods to determine which is most effective in minimizing two kinds of errors of classification.

Both the Bayesian and the Rasch methods described involve complex computer programs and theories of test scores which are unfamiliar to most test practitioners, yet both methods studied showed promise in minimizing the most serious types of classification errors. However, these procedures are both difficult and expensive to implement. Further, the gains made in precision may be offset by the high cost of using either method.

Results using generalizability theory were disappointing. Since the theory is an improvement over classical theory because errors of measurement are defined more precisely, one might expect sizable improvements in decision making. The procedures, as implemented in this study, showed that while generalizability coefficients varied as a function of the conditions studied, standard errors of measurement which were used to set uncertainty bands were not subject to such variation. As a consequence, generalizability theory failed to show superiority in minimizing errors of classification. The binomial and classical approaches also were not effective in reducing errors. Choice of a method for decision making should be governed by which set of criteria are to be employed. These criteria may include familiarity, simplicity, effectiveness in reducing the most serious of classification errors (incorrect classifications), or effectiveness in reducing the least serious of classification errors (uncertain classifications).

One proviso needs to be inserted at this point. The Rasch model is based on a selection of items through empirical means. Items which are matched to the performance level of students are said to give the most

information and thus minimize errors of measurement. In this study, tests were not built in this manner but were created by random sampling. Thus, one needs to build tests using principles of latent trait theory and compare these results to those obtained in the present study.

Regarding the precision of uncertainty bands, setting the bands at 1.0 standard errors, as done in this study, made the uncertainty bands much too wide and seriously minimized the first type of error, incorrect classification. It seems sensible to set smaller bands, perhaps at .75 or .50 standard errors which will have the effect of increasing the first type of error but seriously reducing uncertain classifications. This has the effect of reducing the number of students to receive remedial instruction but increases the number of students misclassified. The consequences of setting different uncertainty bands has not yet been studied.

The most powerful effect observed in this study was the distribution of the student scores relative to passing standards and test lengths. These interactions indicate that where the preponderance of student scores lie with respect to passing standards has much to do with classification accuracy. This need not be an empirical finding, as logical analysis would lead to the same conclusion. However, this study points to the fact that distribution of test scores interacts with other factors to produce conditions where classification accuracy is particularly high or low. Thus, test users are cautioned to anticipate where student scores are likely to fall in a test and plan accordingly. If most student scores fall near the passing standards, many students will be placed in remediation and some will be misclassified seriously. The indication is that instruction should continue to the point that the expectation is that most students' scores will surpass the uncertainty band, thus reducing chances for serious errors (incorrect classifications)

and less serious errors (uncertain classifications). This is an instructional rather than statistical strategy.

Another factor in this study was test length. We know that longer tests are more reliable, and claims that the Bayesian decision-theoretic approach to testing allows for the use of shorter tests was somewhat borne out. However, in general, 10-item tests performed poorly in contrast to longer test lengths studied. These 10-item tests produce large uncertainty bands which trap the greatest proportion of examinees and would result in more extensive remediation. When criterion levels are high, say .80 or more, the bands often exceed 100%, making it impossible for a student to be classified as a pass. In most instances, 15-20 items would appear to be the minimum length for a criterion-referenced test which is used for decision making, and longer tests, up to 30 items, should be used whenever possible. In most cases, improvements in accuracy between 30-item and 40-item tests were small.

Where the criterion level was set was a factor in this study. However, as pointed out earlier, setting criterion levels needs to be based on issues of validity, content or construct. It is clear that we can manipulate criterion levels to achieve greater classification consistency, but this seems inappropriate in light of the argument that criterion levels are set to qualify successful students for future work. Adjusting the criterion levels to reduce classification errors runs counter to the procedures one uses to set standards to achieve content valid or construct valid interpretations of test scores.

While this study contrasts the most prominent methods for studying the reliability of test scores in systematic instruction, it has omitted several approaches due to the insufficiency of specific methods of application.

Further, the latent trait models have not been fully or adequately tested in comparison to the other methods. Future studies are needed which empirically contrast these differing models in light of their consequences on the two types of classification errors. Finally, models must be studied with respect to their relative merits in the areas of practicality, cost, efficiency and acceptability. Any analysis in these latter areas would favor the least restrictive models, such as the binomial, but this kind of analysis has not been done.

Reference Notes

1. Haladyna, T. An analysis of two procedures for decisionmaking when using domain-referenced tests. Paper presented at the annual meeting of the National Council for Measurement in Education, Washington, D.C., March-April 1975.

2. Shaycroft, M. F. A paradox in setting cutting scores on criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

3. Shaycroft, M. F. The coefficient of accuracy: A new statistic for criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

4. Bernknopf, S., & Bashaw, W. L. An investigation of criterion-referenced tests under different conditions of sample variability and item homogeneity. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

5. Epstein, K. I., & Knerr, C. S. Criterion-referenced test interpretation of "classical" measurement theory. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

6. Haladyna, T. M. An investigation of full and subscale reliabilities of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

7. Moyer, J. E., & Fishbein, R. L. A comparison of Kuder-Richardson Formula 20 and Kappa as estimates of the reliability of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

8. Haladyna, T., & Roid, G. R. An empirical comparison of three approaches to achievement testing. Paper presented at the annual meeting of the American Psychological Association, San Francisco, 1977.

9. Frederick, B. A(nother) model for the reliability of criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

10. Huynh, H. The kappamax reliability index for decisions in domain-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

11. Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

References

Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1, 1-12.

Brennan, R. L. Extensions of generalizability theory to domain-referenced testing (ACT Technical Bulletin No. 30). Iowa City: The American College Testing Program, 1978.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: John Wiley, 1972.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 11, 371-400.

Hambleton, R. K., & Cook, L. L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objective-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-190.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harris, C. W., Pearlman, A. P., & Wilcox, R. R. Achievement test items--methods of study. Los Angeles: Center for the Study of Evaluation, UCLA, 1977.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264. (a)

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 4, 65-78. (b)

Keller, F. S. Goodbye, teacher.... Journal of Applied Behavior Analysis, 1968, 1, 79-89.

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement, 1972, 9(1), p. 31.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Meskauskas, J. A. Evaluation models for criterion-referenced tests: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Messick, S. The standard problem: The meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J. Passing scores and test lengths for domain-referenced tests. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Company, 1974. (a)

Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14, 17-21. (b)

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38(1), 19-46.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Nunnally, J. Psychometric theory. New York: McGraw-Hill, 1967.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, inc., 1978.

Roid, G. H. (Ed.), Dental anatomy: A self test instructional program, Vols. 1-5, Revised Edition. Monmouth, Oregon: Teaching Research, 1977.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Wright, B. D. Sample-free test calibration and person measurement. Invitational Conference on Testing Problems. Princeton, N. J.: Educational Testing Service, 1967.